Administrative Data Based Population Estimates Scotland 2016–2018: Methodology Report (v2)

Statistical Research

Published on 14 December 2021

Disclaimer: These statistical research outputs are **not the OFFICIAL STATISTICS** for Population Estimates for Scotland. The Official Statistics can be found at the statistics and data section of National Records of Scotland's website.

This publication reports on the results of research into how population estimates might be produced using a range of administrative data.

Any presentation or use of these research outputs should make clear to users the nature and purpose of the statistics.

Contents

1. Abstract

National Statistics for population estimates in Scotland are provided by the census and the mid-year population estimates (both published by National Records of Scotland (NRS)). NRS is now exploring the possibility of using de-identified data from public-sector administrative datasets to produce population estimates. The methodology used to produce the estimates presented in this publication is broadly similar to the methodology used for the previously-published 2016 administrative-data based population estimates (v1).

The change with the largest impact on the estimates is that the business rule for inclusion only considers records from the Health Activity dataset if the most-recent interaction with the health service is within a certain time of the reference date. The period considered is age dependent. Ages at which persons tend to have more interactions (such as older persons) have shorter periods. For example, if a person over 50 had last interacted with the NHS two years prior to the reference date, then that would not be taken as evidence that they are in the Scottish population. This change reduces the estimates, in many cases bringing them slightly closer to the official mid-year estimates.

The business rules are applied to a dataset where personal information has been de-identified, so the statistician does not know any names, dates of birth or addresses.

Other significant changes from v1 are:

- Including persons with an armed forces posting on the NHSCR, even if they do not appear on any other dataset
- Excluding persons whose location cannot be identified
- Moving or excluding persons who appear in data zones known to be empty on the reference date
- Reassigning prison populations to the data zone where the corresponding prison is located
- Not considering higher and further education records for the academic year following the reference date, but only for the academic year immediately prior to the reference date.

2. Method Summary

The following is an outline of the steps involved in the method for producing the Administrative Data Based Population Estimates (ABPE). These steps are discussed in more detail in Section 4 of the methodology report for the 2016 publication. The methodology used for the publication of the 2016 data will be referred to as version 1 (v1). The methodology used for the current publication (presenting revised 2016 estimates, along with estimates for 2017 and 2018) will be referred to as version 2 (v2). This document will focus more on the changes to the methodology from v1 to v2. These changes do not affect the broad principles of the v1 methodology, but rather are refinements to it.

Broadly, the method to produce the ABPE for each year was as follows:

1. Standardise the identifiable data
2. Separate payload (for example, age, sex) and identifiable data (name, date of birth and postcode)
3. Generate linking variables
    a. Name variables (for example nicknames, first few letters of last name)
    b. Postcode levels (postcode unit, sector, district and area)
    c. Date of birth Bloom filters[1]
4. De-identify identifiable data
5. Transfer data to the National Safe Haven
6. Join payload and de-identified datasets
7. Concatenate datasets and link between and within the datasets using the linking variables
8. Trim links to remove links that link pairs of records that likely represent distinct individuals
9. Resolve linked records to individuals, assign UPIDs[2], and produce the ADRS[3]
10. Apply business rules to remove persons believed not to be present in the population at the reference date, and produce the SIDD[4]
11. Assign analysis variables (for example, council area)

Linking multiple datasets together could potentially increase their sensitivity. Therefore, the datasets to be linked are only brought together in the same processing environment once they have been de-identified. In order to maintain high levels of security, and preservation of privacy, the process is divided into three parts. Each part processed in different environments, and by different individuals. In this way, none of the team (apart from the NRS Head of Admin Data) see the identifiable data of multiple datasets simultaneously. For more detail see the Data Protection Impact Assessment (DPIA) and Section 2 of Administrative Data Based Population Estimates, Scotland 2016: Methodology Report (v1).

[1] See Section 4.4.2 of ABPE, Scotland 2016 Methodology Report (v1) for more detail on Bloom filters.
[2] Unique Person IDs.
[3] Administrative Data Record Set.
[4] Scotland's Integrated Demographic Dataset.

3. Datasets Used

A range of administrative datasets are provided by suppliers. These are specified in Table 1. The datasets used for the revised 2016 ABPE are the same as those used for the v1 publication. The datasets used for 2017 and 2018 are similar and have the same sources as the equivalents for 2016.

Table 1 Datasets used along with the supplier that provided them and a brief description.

Name	Source[5]	Description
NHSCR	NRS	Persons born or died in Scotland or registered with a Scottish GP.
Health Activity	PHS	Health activity data, last interaction with Health Service but does not include any medical records.
Electoral Register	EROs	Persons registered to vote, including for Scottish elections (16–17 year olds). Dates of birth are not supplied.
Scottish Pupil Census	SGLD	Children and young persons enrolled at publically funded schools in Scotland. Names are not supplied.
Vital Events	NRS	Births, deaths and marriages registered in Scotland, including the details of the parents of children born and those registering deaths.
HESA	HESA	All students studying at Scottish higher education providers and Scottish domiciled students studying at higher education providers in England, Wales and Northern Ireland.
FES	SFC	Students studying in colleges in Scotland, except those studying in Higher Education programmes in the University of Highlands and Islands or Scotland's Rural College.
Geography	NRS	Lookups from de-identified postcode to council area, data zone, SIMD and urban–rural classifications.

More details on these datasets can be found in:

- Section 3 of ABPE Scotland 2016: Methodology Report (v1)
- Quality Assurance of Administrative Data 2016
- Quality Assurance of Administrative Data 2017 and 2018

[5] See Glossary in Annex 1 for expansion of abbreviations.

4. Differences Between Methodology v1 and v2

4.1 Data Zone–Level Adjustments

Part of the investigation to understand the observed differences between the previously released ABPE and the published mid-year estimates was to analyse differences at data zone level[6]. When this was done it was found that there were large differences for a small number of data zones. Investigation of these individual data zones revealed some patterns, and adjustments were made accordingly. At the moment, data zone information has not been published as we have not been able to assess its quality over the three years, but it can help with general business rule investigation.

Prisons

It was found that many of the data zones that are known to contain prisons had notably fewer persons than published estimates for the sex that the particular prison accommodated[7]. There was also correspondingly more persons in a nearby data zone, where there was no known large communal establishment such as prisons, care homes or student halls. It appeared likely that on the health datasets prisoners were often being recorded at a nearby health centre, rather than at the prison itself.

To correct this, the postcode in the data zone with an unusually high number of persons of the relevant age distribution and sex was identified. The persons of the relevant sex were then moved from that postcode to the data zone that contained the prison. Note that postcodes cannot be identified as the postcode information has been de-identified, but it remains possible to group people together if they are in the same de-identified postcode. There may be some extra persons who are not in prisons being moved by this adjustment, but it was felt that is better for the overall methodology. Recall again that as the data have been de-identified it is not possible to identify individuals who are in prison.

Vacant Data Zones

Intelligence available to the NRS demographic statistics team, provided by local authorities, identifies some data zones that have zero population. This can happen where an area with high-density housing has that housing demolished and not immediately replaced at that location. Over time this could lead to a redrawing of data zone boundaries to ensure that data zones have a roughly equal population.

---

[6] Data zones are the key geography for the dissemination of small area statistics in Scotland and are widely used across the public and private sector. Composed of aggregates of Census Output Areas, data zones are large enough that statistics can be presented accurately without fear of disclosure and yet small enough that they can be used to represent communities. They are designed to have roughly standard populations of 500 to 1,000.

[7] The locations of prisons in Scotland is available at https://www.sps.gov.uk/Corporate/Prisons/Prisons.aspx, along with information on whether they hold male or female prisoners.

However, using the 2011 data zone boundaries means that adjustments need to be made for the relevant data zones.

The data zones in question are S01010206, S01010226 and S01010227. If a record was placed in one of those data zones then an alternative postcode was sought for the record. For example NHSCR might place the person in S01010206, but if the person appeared on the electoral register at a different de-identified postcode then that de-identified postcode was used instead. If there were no alternative de-identified postcodes available then the person was removed from the dataset.

These data zones could also provide information about how up to date and reliable the location information on each dataset was. Prioritising the location of persons on their NHSCR record resulted in more persons in these vacant data zones than when prioritizing the Health Activity datasets. In general the Health Activity locations were more reliable than those on NHSCR (although it should be noted that in most cases the de-identified postcodes agreed between the two datasets). Therefore the prioritisation was changed so that the Health Activity dataset was given the highest priority for all records.

4.2 Health Activity Interaction Thresholds

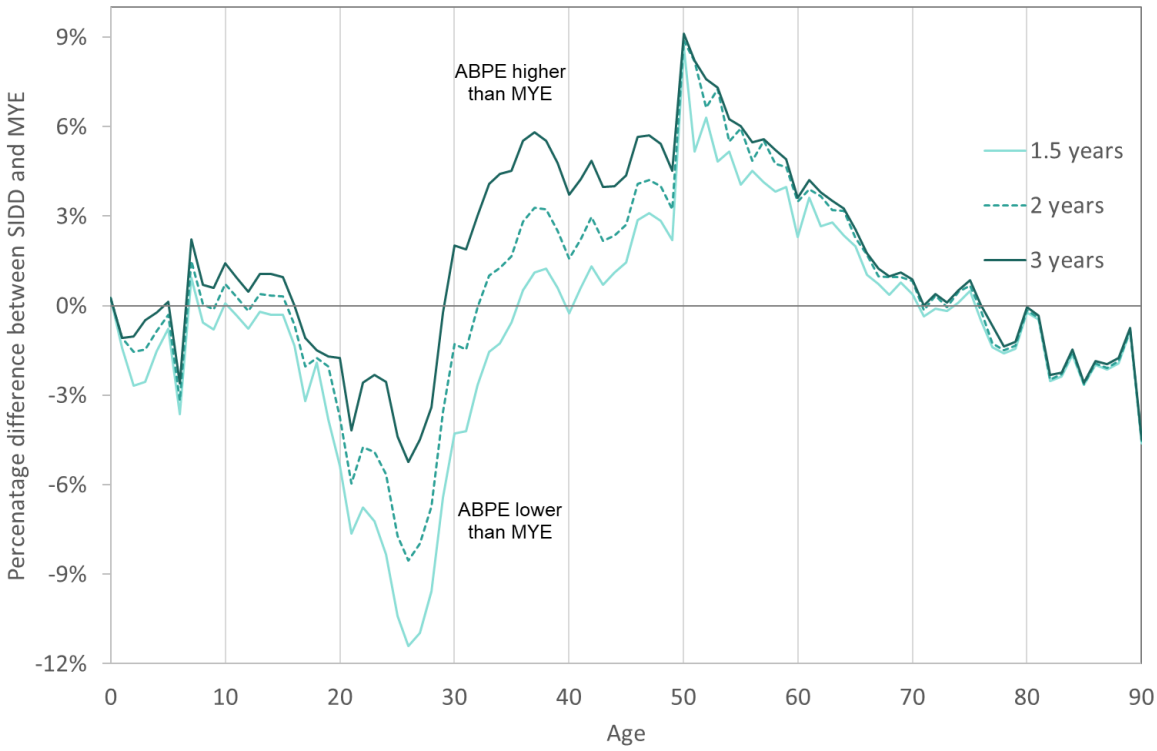
For the Health Activity datasets, information is now available for each record to indicate when the person last interacted with the NHS (as at the dataset reference date). Note that the Health Activity datasets give no indication of what the interaction was about, only the date it occurred; there is no access to medical records in these datasets. The interaction times can be used to indicate the confidence that can be placed in a particular record. For example if someone had very recently interacted with the NHS, then it is likely that their information is correct on the reference date. However, if there had been no observed activity for a person then this could raise doubts as to whether they were still at the recorded location.

Reducing the threshold for which Health Activity records are included can also help deal with over-coverage. Figure 1 shows how the 2017 SIDD would compare with the MYE for males for different Health Activity thresholds. (The SIDD is Scotland's Integrated Demographic Dataset. This is the dataset of records after the business rules have been applied. The ABPE are based on the SIDD.) For example, the 2-years series indicates what the difference would be if the SIDD only considered records on the Health Activity dataset where the latest interaction was after 30/6/2015.

This shows that over-coverage is most problematic for persons (male in this case) in their 50s, but that we have under-coverage for people around their 20s. The equivalent figures for other years, and for females, are shown in Annex 2 for completeness. These show broadly similar patterns over the three years. It would be possible to choose a threshold for each age that results in estimates that are closest to the mid-year estimate. However, such fine tuning of the model could not be independently justified, and would be unlikely to be stable from year to year.

Whatever thresholds are used should be guided by, but be independent of, the MYE. This will allow the resultant SIDD to be tested against the Census 2022 to measure the validity and reliability of the ABPE.

Figure 1 Potential difference between the SIDD and the MYE for males in 2017 for different cut-offs of the Health Activity interaction date.



An alternative method of justifying age-specific rules is desirable. One way to do this would be using the distribution of interaction dates. The following figures consider persons who appear on the Health Activity dataset, and also meet other requirements for inclusion on the SIDD. That is, they appear on the NHSCR and have none of the conditions for exclusion, such as having a death registered. This covers the vast majority of persons who appear on the Health Activity dataset, but excludes anyone who would not be considered for the SIDD. If thresholds based on timing of last interaction are incorporated, then some of those persons would no longer be included in the SIDD. For the purposes of this discussion, the SIDD that would be produced if there were no conditions on interaction dates is referred to as SIDD+. Persons who appear on the NHSCR and Health Activity (but no other datasets) and have their latest interaction date before the interaction threshold would therefore appear on the SIDD+, but not the SIDD. Therefore the set of persons on the SIDD is a subset of persons on the SIDD+.

Figure 2 shows the proportion of females on SIDD+ with a Health Activity interaction date more than two years prior to reference date (30 June), for 2016 (revised), 2017 and 2018. Figure 3 shows the equivalent for males. Recall that any person without an interaction in the previous three years would not be included in the Health Activity dataset.

Figure 2 The proportion of female Health Activity records on SIDD+ that have the last interaction date more than 2 years before the reference date.

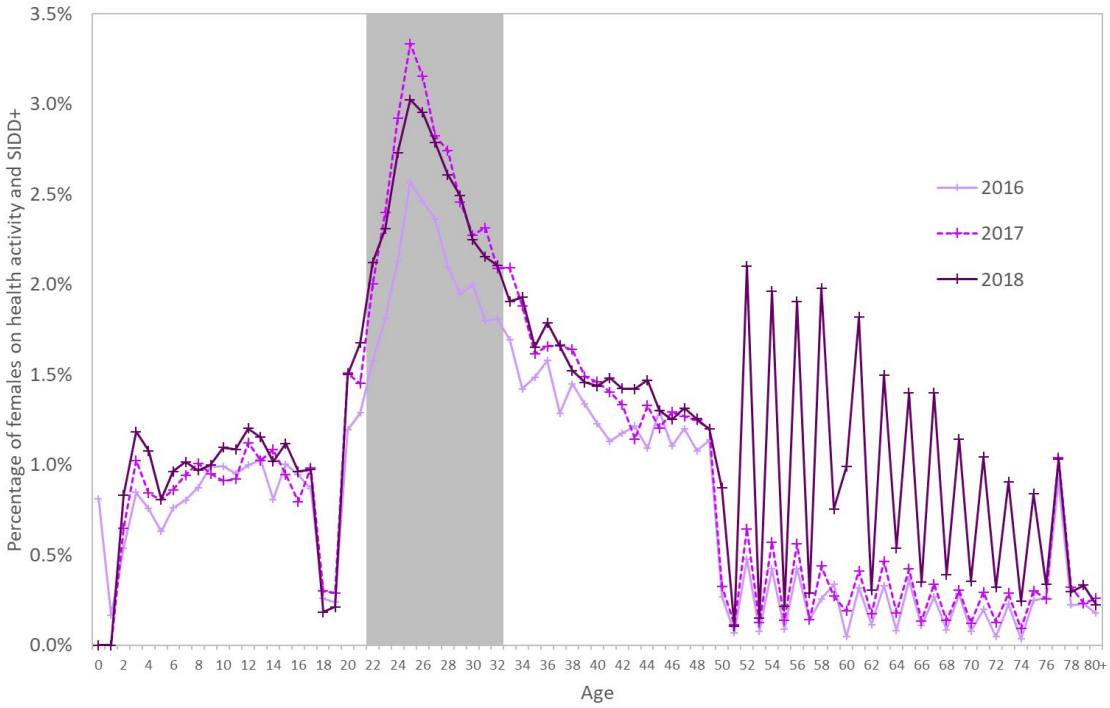
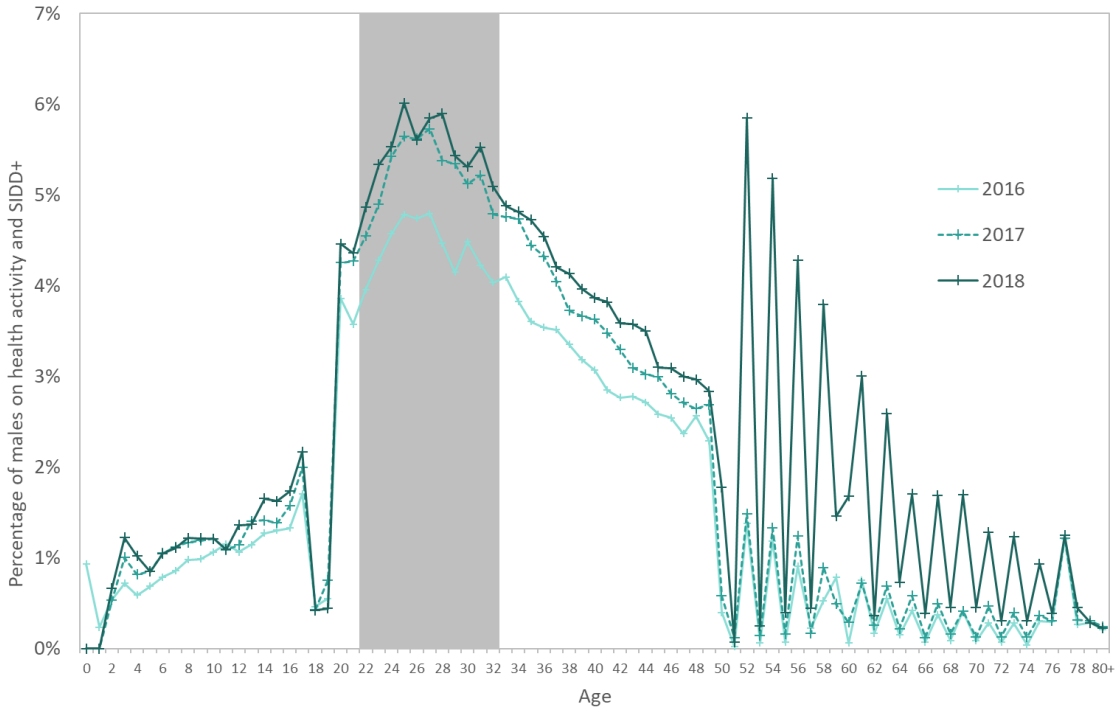


Figure 3 The proportion of male Health Activity records on SIDD+ that have the last interaction date more than 2 years before the reference date.



At ages when most persons have recent interactions, it would seem more likely that persons who have not had interactions for a long time are in fact no longer living at

the recorded location. Conversely if a high proportion of people have not had an interaction for a long period then it seems less likely that the lack of a recent interaction can be taken as evidence that they have moved away.

The ages where a high proportion of people have interactions a long time before the reference date should be the ages where the longest interaction times are considered for inclusion. Figure 2 and Figure 3 consider the SIDD+. The figures show that ages 22–32 (shaded grey in the figures) have among the highest proportion of people with last interaction more than two years before reference date. The exact distribution varies by year and sex, but across the three collection years this age range tends to capture the highest proportion. This seems a plausible age range where persons would have less need to interact with the health service. Therefore, for persons aged 22–32, the business rule will include records from the Health Activity dataset where the last interaction is within three years of the reference date. That is the only age group where the threshold is three years.

Note that figures 2 to 5 show that higher proportions of 18 year olds have recent interactions than most other ages. The reasons for this are not fully clear. However, it is known that students under the age of 25 attending university for the first time were, from Autumn 2015, offered the MenACWY (meningitis) vaccine[8], which may have an impact on this.

Older persons would be among those expected to interact with the health service most often. For such persons it may be desirable to exclude records from the Health Activity dataset if they did not have recent activity. Figure 2 and Figure 3 show that from age 50 onwards comparatively few persons have interactions more than two years before the reference date, especially for 2016 and 2017. It can be seen that there is an approximate two-year pattern in activity from this age onwards. This is because there is a bowel cancer screening programme that happens every two years for those aged 50 and over. Figure 4 and Figure 5 show the equivalent to Figure 2 and Figure 3, but show the proportion of people with last interaction more than 1.5 years prior to the reference date. The business rule will therefore not include persons aged 50 and over if their last interaction on the Health Activity dataset was more than 1.5 years before the reference date.

Note that the pattern among those aged over 50 is notably different in 2018 than in the other years. The reasons for this are not fully understood. When the 2019 data is available it will be analysed to explore whether the 2018 pattern persists or whether it is unique to 2018. The results of that could give clues to the origin of the difference. It is known[9] that in November 2017 the test for the bowel screening programme changed, and that this affected the uptake for the programme.

[8] See www.hps.scot.nhs.uk/a-to-z-of-topics/meningococcal-disease.
[9] See https://beta.isdscotland.org/find-publications-and-data/conditions-and-diseases/cancer/scottish-bowel-screening-programme-statistics/4-february-2020.

Figure 4 The proportion of female Health Activity records on SIDD+ that have the last interaction date more than 1.5 years before the reference date.

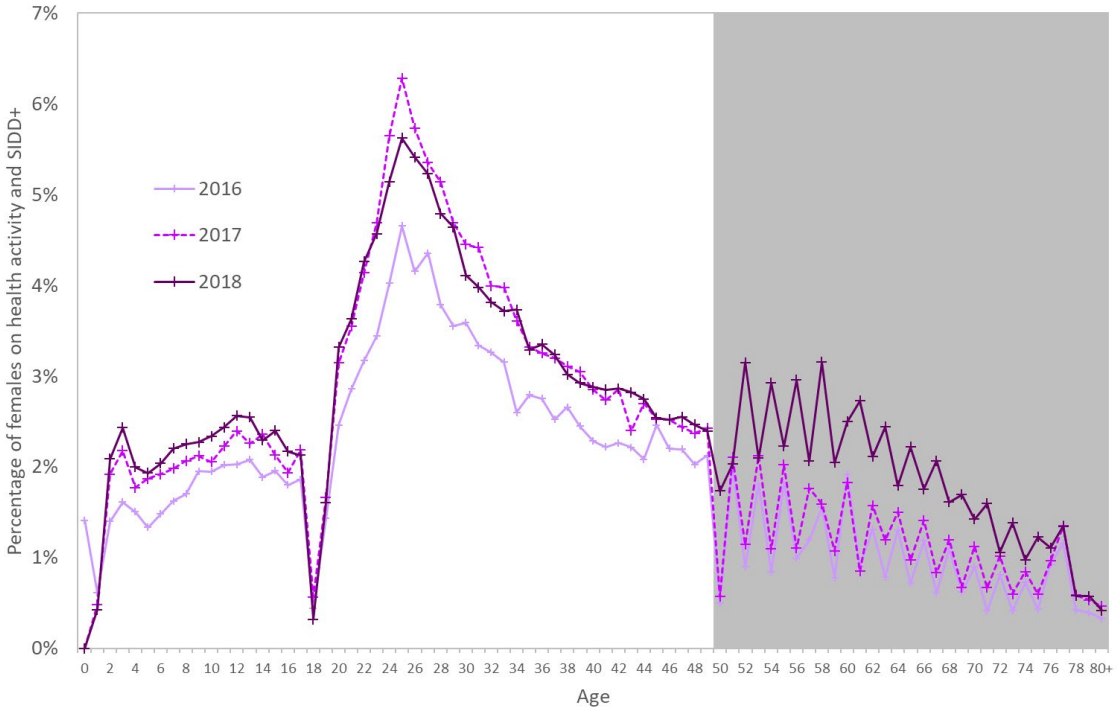
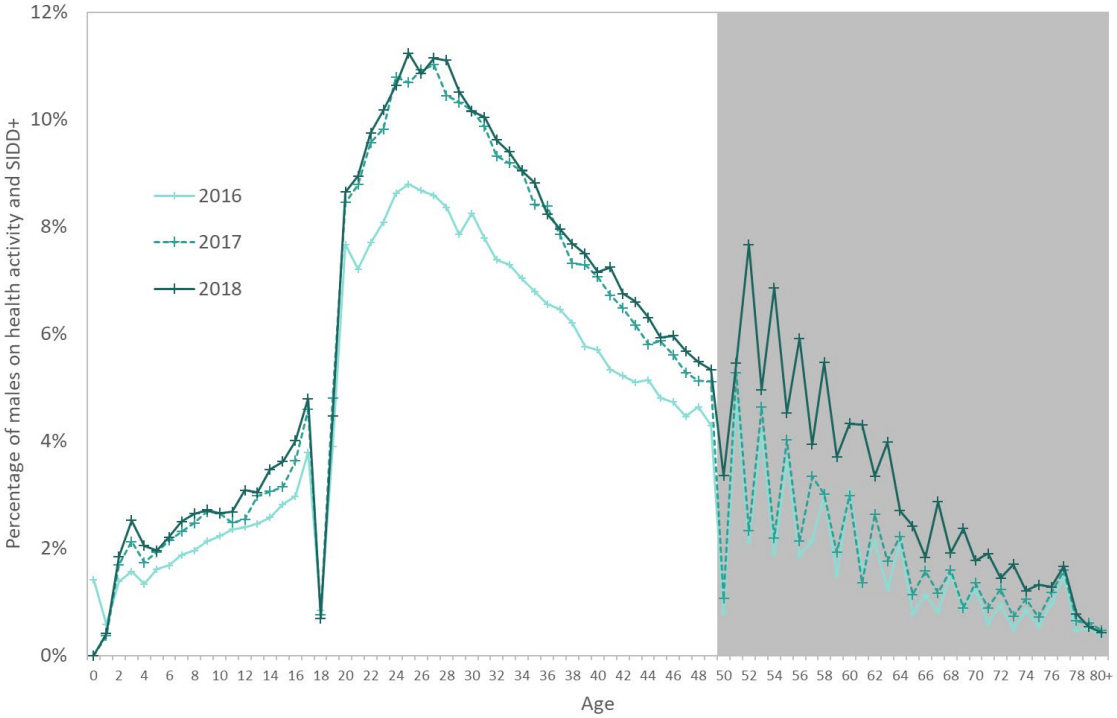


Figure 5 The proportion of male Health Activity records on SIDD+ that have the last interaction date more than 1.5 years before the reference date.



Given the explainable observed two-year pattern of activity for persons aged 50 and over, it had been hoped that the business rule could reflect this by having a different rule for persons with odd ages as for those with even ages. However, it became apparent that this approach would not be successful. Firstly, the pattern does not appear to have an exact 2-year pattern. In Figure 2 and Figure 3 persons in their 50s tend to have peaks for even ages, while by the 60s this tends to be for odd years. Secondly, the pattern is not consistent across years, with 2018 tending to have the opposite pattern to 2016 and 2017 in Figure 4 and Figure 5.

Table 2 summarises the thresholds used for the different age ranges.

Table 2 Threshold for the interaction date on the Health Activity dataset to be considered for inclusion on the SIDD by age.

Age range	Threshold for inclusion
0–21	2 years
22–32	3 years
33–49	2 years
50+	1.5 years

4.3 Other Changes to Business Rules

Excluding Records with Missing Location

In the v1 methodology UPIDs were excluded from the SIDD if they had missing sex or age information. However, missing location information did not exclude them from the SIDD. UPIDs could have missing location information if a de-identified postcode was not recorded for the person on any of the datasets that contributed to it. It could also happen if the de-identified postcode that was assigned to the UPID was not included in the geography lookup. That may be because it was an invalid or out of date postcode.

For v2 a missing or invalid location is taken as evidence that the information relating to the person is not reliable. Such records are therefore excluded from the SIDD. This also has the benefit that the final dataset is complete in terms of location, as well as age and sex. In v1 for 2016, 0.12% of the records on the SIDD had missing location.

HESA and FES from Academic Year Just Finished

In v1 two HESA datasets were included in the datasets that were linked together. Presence on either was taken as evidence of being in the population, but the location was prioritised for the permanent, rather than term-time address. This approach was taken as it was felt that the reference date of 30 June fell outside the academic year, and so students would not be at their term address.

However, discussions with NRS demography team revealed that the published mid-year estimates try to reflect students at their term-time address. This is felt to more accurately represent where persons reside for most of the year. Therefore students would be placed at where they were expected to be during term time for the academic year that had just finished.

In order to align with this approach, v2 only uses one dataset for HESA, using the one for the academic year just finished. For example for the 2016 estimates the HESA dataset covering the 2015/16 academic year was used. The same approach was used for the FES dataset.

Ministry of Defence Bases

Data zone-level comparisons of the ABPE and the MYE revealed that some of the data zones with large differences in numbers included a known Ministry of Defence (MoD) base. For these data zones the ABPE was generally found to be lower than the MYE equivalent, so measures were taken to increase the number of armed forces personnel in the ABPE. As with all the datasets used, the NHSCR is de-identified so we cannot identify individuals by name, date of birth or postcode, but it does hold some additional posting information against each record. The v1 ABPE utilized the NHSCR posting variable to identify persons who were believed to have left Scotland, or to have died (in addition to death registrations). The posting variable also indicates whether the de-identified person is a member of the armed forces (and part of the MoD health service). Normally, to include a de-identified person on the SIDD they need to appear on the NHSCR and at least one other dataset. For persons who appeared on NHSCR with an armed forces posting, dropping the condition that persons should appear on another dataset resulted in estimates that were closer to the MYE for data zones containing known MoD bases. This change was therefore made to the business rules for v2.

4.4 Health Activity 2017 Postcode Corrections

After linking, an issue with some of the de-identified postcodes was discovered. By linking NHSCR and Health Activity records exactly on de-identified name and date of birth it was found that many records had different de-identified postcodes; Table 3 shows there are substantially more in 2017 than in 2016 or 2018. This issue had not been apparent when the datasets were quality assured in isolation as the postcodes were generally valid.

For simplicity, this analysis was done on a version of the Health Activity datasets that had combined the secondary and primary care datasets. Individual persons may appear on only the primary care dataset, only the secondary dataset, or appear on both. Therefore the combined Health Activity dataset contains more individuals than either the primary or secondary care datasets. Not all those individuals could link exactly to a NHSCR record, and so the total of Table 3 is smaller than the total number of individuals on the Health Activity datasets.

Table 3 Comparison of NHSCR and Health Activity de-identified postcodes.

NHSCR–Health Activity postcode comparison	2016	2017	2018
Same	5,317,347	3,129,767	5,429,618
Different	222,351	2,539,724	186,212
Total	5,539,698	5,669,491	5,615,830

The records where the de-identified postcodes differed between the two datasets were then linked to the marriages dataset. The postcodes on the marriages dataset had much stronger agreement with the NHSCR than with the Health Activity dataset (see Table 4), so the issue with the postcodes was probably on the (2017) Health Activity dataset.

Table 4 Comparison of 2017 NHSCR, Health Activity and marriages de-identified postcodes, where different between NHSCR and Health Activity.

Postcode comparison	Number of records
Marriages, NHSCR and Health Activity all different	6,022
Marriages agrees only with Health Activity	981
Marriages agrees only with NHSCR	18,901
Total	25,904

To address this issue all the records from NHSCR and Health Activity datasets for 2016, 2017 and 2018 were linked together using exact linking on de-identified name and date of birth. Cases where a person appeared on all six datasets were identified. Cases where the combination of de-identified name and date of birth was not unique were removed.

Table 5 Number of changes made to the (de-identified) postcodes on the 2017 primary and secondary care Health Activity datasets.

Dataset	Unchanged	Changed	Total
Primary Care	4,192,447	1,533,967	5,726,414
Secondary Care	2,771,505	1,020,848	3,792,353

The de-identified postcodes of those remaining were then compared. If the same de-identified postcode appeared on all three NHSCR datasets and the Health Activity datasets for 2016 and 2018, then the de-identified postcode on the Health Activity 2017 dataset was changed to the postcode that appeared on the other datasets. This corrected many of the postcodes, but because many persons change postcode over the 2016 to 2017 period not all errors could be confidently corrected (see Table 5). These corrections were made to the original primary and secondary care Health Activity datasets that were used in the linking. The large number of cases where the

NHSCR 2017 de-identified postcode agrees with the 2016 and 2018 NHSCR and Health Activity postcodes gives further evidence that it is the Health Activity postcodes that should be changed.

Having multiple datasets linked together can help mitigate problems caused by the remaining problematic de-identified postcodes. For example, suppose a person appears on the NHSCR, Health Activity and Electoral Register (ER) datasets. If they appeared at the same postcode in NHSCR and ER, then these two records would be assigned the same UPID, while the Health Activity record would get a different UPID. In such cases the UPID with the ER would be included in the SIDD, more than likely at the correct location. The UPID for the Health Activity record would be excluded from the SIDD as it did not link to an NHSCR record.

This issue was not fully mitigated against however, because 2017 was also the year that Fife ER data was not included due to a clerical error with passwords. This means that persons in Fife are more likely to appear on just the NHSCR and Health Activity datasets than persons in other authorities. As the location is prioritised by the location on the Health Activity dataset, this means that persons who should be located in Fife are more likely to be placed at the wrong location. This results in estimates for Fife being further below the mid-year estimates in 2017 than for other years, and lower than many (but not all) other authorities. (More information can be found in the Quality Assurance of Administrative Data, 2017–2018.)

4.5 Business Rules for 2016 (revised), 2017 and 2018

The business rules for UPIDs to be excluded from the SIDD (even if the inclusion conditions are met) are similar in v1 and v2. Specifically, if any of the following conditions are met then the UPID is excluded from the SIDD for v1 and v2:

- UPID appears on the death registrations with a death date on or before the reference date
- UPID appears on the NHSCR with a posting indicating that the person has died, has moved elsewhere in the UK, or outwith Scotland (embark)
- UPID appears on one of the electoral register datasets where the franchise is ‘F’, indicating that the voter lives overseas
- No information for age or sex is available for the UPID.

The following exclusion conditions apply to v2, but did not apply to v1:

- No information for location is available for the UPID
- The UPID is placed in a data zone known to be empty, and no alternative location is available.

If none of the above conditions are met then the UPID will be included on the SIDD if at least one of the inclusion conditions are met. These differ between v1 and v2. The inclusion conditions for v1 and v2 are indicated in Table 6.

Table 6 Inclusion conditions for ABPE v1 and v2.

V1	V2
UPID appears on the birth registrations as a child and is aged below one.	UPID appears on the birth registrations as a child and is aged below one.
UPID appears on NHSCR with a Scottish posting, and at least one other dataset.	UPID appears on NHSCR with a Scottish posting, and at least one other dataset. The UPID is only counted as appearing on the Health Activity dataset if the latest interaction is within a certain time of the reference date (see Table 2 for details). UPID appears on NHSCR with an armed-forces posting.

4.6 Comparison of v1 and v2 Results

Figure 6 Differences between the 2016 ABPE and MYE for v1 and v2 of the ABPE by sex and age.

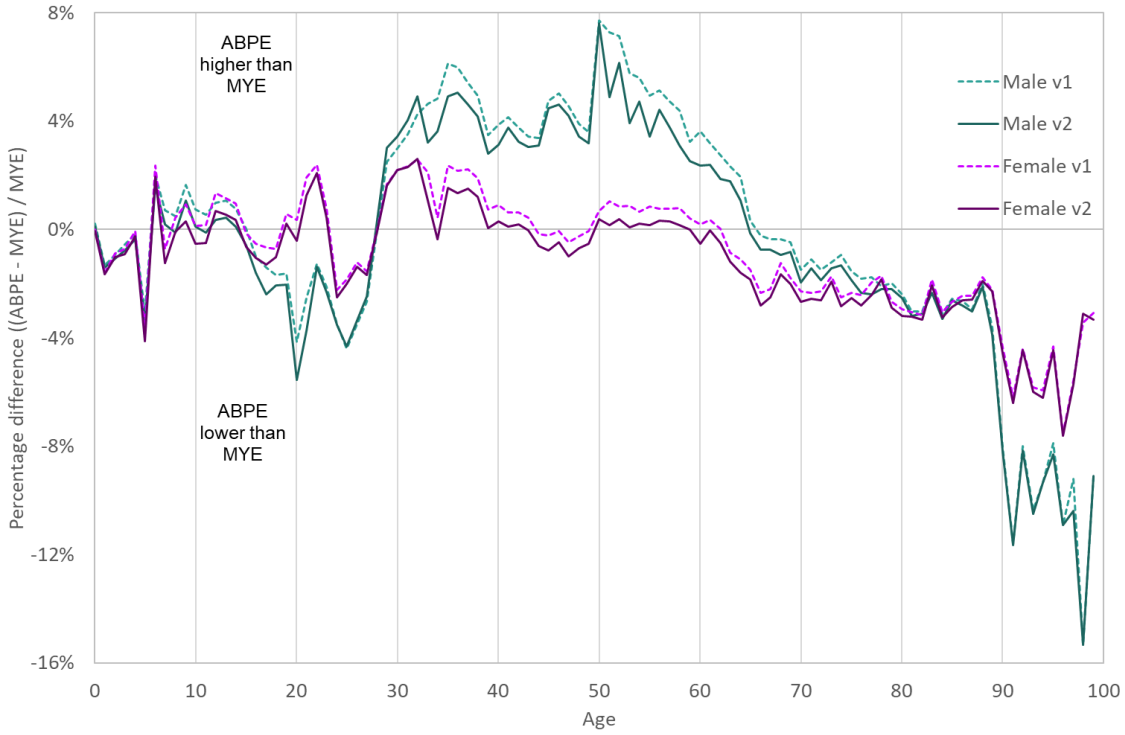


Figure 6 shows the difference between 2016 ABPE v1 and v2 , when compared with the MYE by age and sex. In general the two versions are similar, although v2 tends to have lower estimates then v1. This is primarily because of the introduction of a threshold for the interaction date for the Health Activity dataset. For many ages this brings the estimate closer to the MYE, particularly for females, where the estimates tend to be closer to begin with. For older persons this takes the estimate further from

the MYE, but the differences between the estimates are smaller here, as older persons are more likely to have had recent interactions.

It can also be seen that at some ages v2 is higher than v1, particularly for males around age 30. This will be due to the inclusion of persons with an armed forces posting, even if they do not appear on other datasets. It seems plausible that this affects males more than females, given that in 2016 only 10.2% of UK regular armed forces personnel were female (Ministry of Defence, 2016). At ages 22–32 there is no difference between v1 and v2 in terms of the Health Activity last interaction threshold, which will be why the armed forces effect is most apparent for those ages. Above age 32, v2 drops below v1 where the change to the interaction date threshold will have a larger effect than the armed forces effect.

5. References

Ministry of Defence (2016) *UK Armed Forces Biannual Diversity Statistics 1 April 2016*, [Online] available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/530586/Biannual_Diversity_Statistics_1Apr16_revised.pdf (accessed 27 October 2021)

Annex 1: Glossary

Term	Definition
ADRS	Administrative Data Record Set. Once the source datasets have been linked, the ADRS contains a record for each individual who appears on these source datasets.
De-identified	Converting variable values to a new value in a non-reversible way. This is done in such a way that small changes to the original values are highly likely to result in a very different final value. The process was done using the SHA-256 hashing algorithm.
ER	Electoral Register.
EROs	Electoral Register Officers (www.saa.gov.uk/electoral-registration).
HESA	Higher Education Statistics Agency (www.hesa.ac.uk).
MoD	Ministry of Defence.
MYE	Mid-Year Population Estimates for Scotland (www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates).
NHSCR	National Health Service Central Register. A dataset of persons registered with an NHS GP in Scotland, or who were born or died in Scotland.
NRS	National Records of Scotland (www.nrscotland.gov.uk).
PHS	Public Health Scotland (https://publichealthscotland.scot).
SFC	Scottish Funding Council (www.sfc.ac.uk).
SGLD	Scottish Government Learning Directorate (https://www.gov.scot/collections/school-education-statistics/)
SIDD	Scotland's Integrated Demographic Dataset. This is similar to the ADRS, but only contains records that have passed the business rules (that is, records for which it is believed that the person they represent is present in the population at the reference date).
SIDD+	As SIDD, but without applying a condition on the interaction date for Health Activity.
UPID	Unique Person ID. UPIDs are assigned to administrative data records after linking. Records believed to represent the same person will be assigned the same UPID. Conversely, records believed to represent different persons will have different UPIDs.

Annex 2: Supplementary Figures

Figure 7 Potential difference between the SIDD and the MYE for males in 2016 for different cut-offs of the Health Activity interaction date.

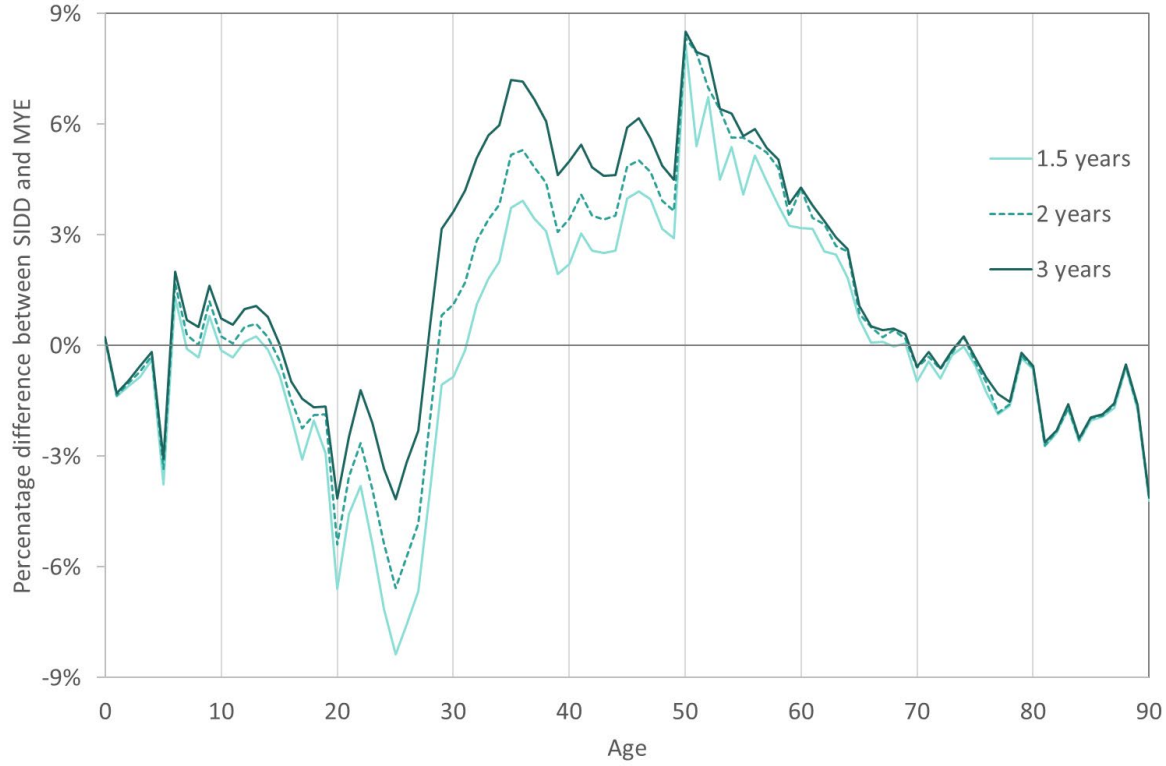


Figure 8 Potential difference between the SIDD and the MYE for males in 2018 for different cut-offs of the Health Activity interaction date.

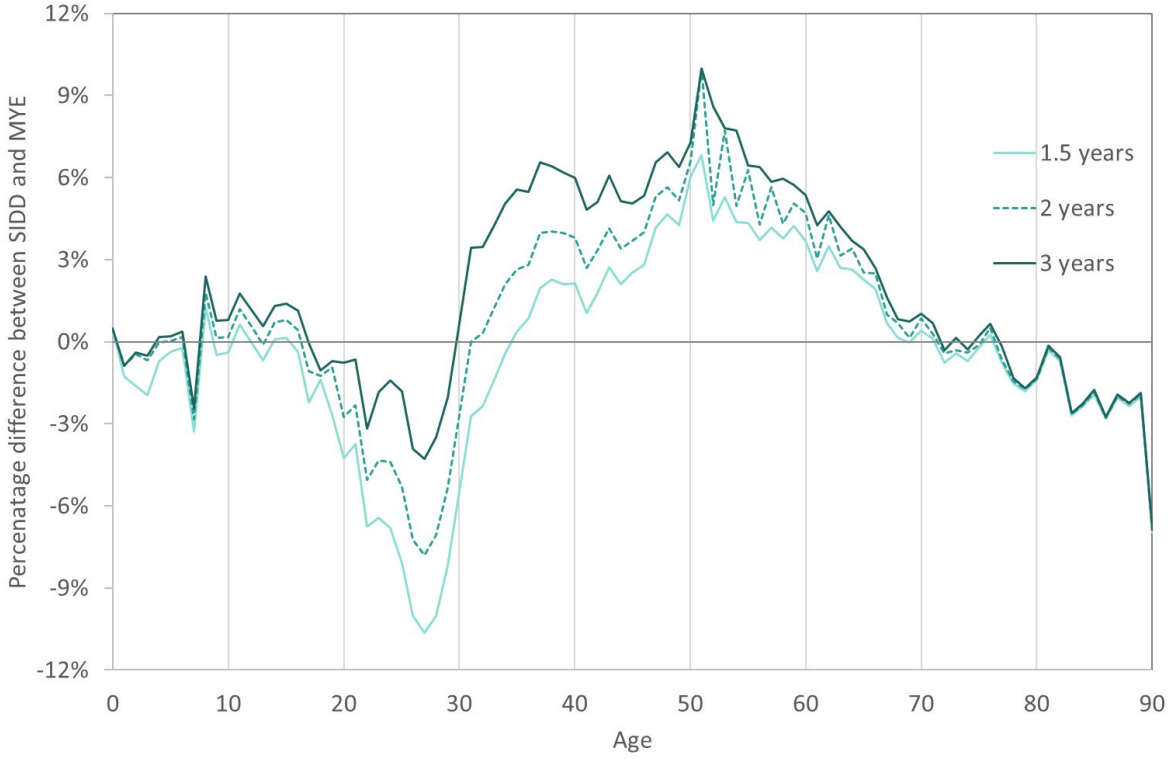


Figure 9 Potential difference between the SIDD and the MYE for females in 2016 for different cut-offs of the Health Activity interaction date.

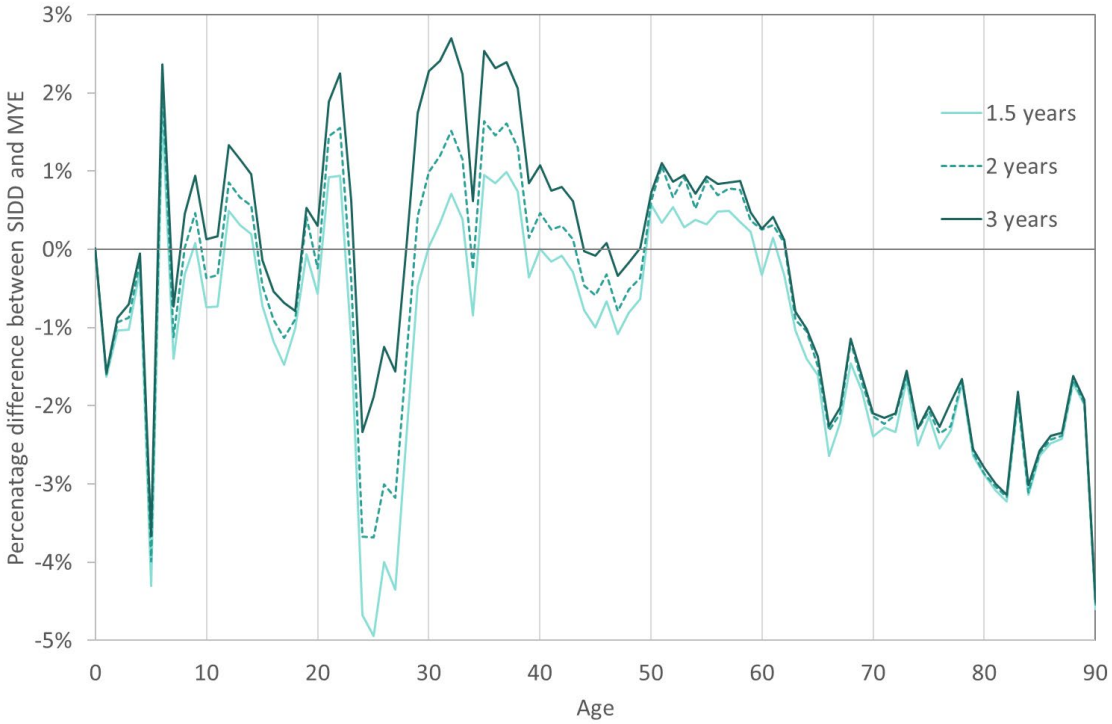


Figure 10 Potential difference between the SIDD and the MYE for females in 2017 for different cut-offs of the Health Activity interaction date.

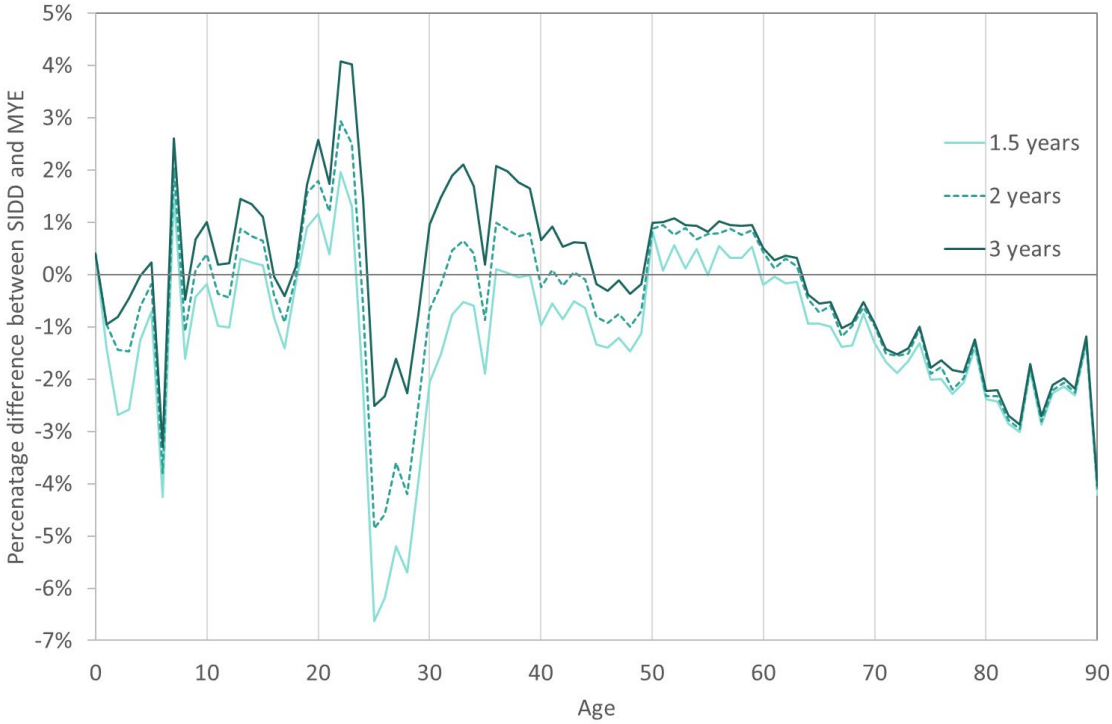
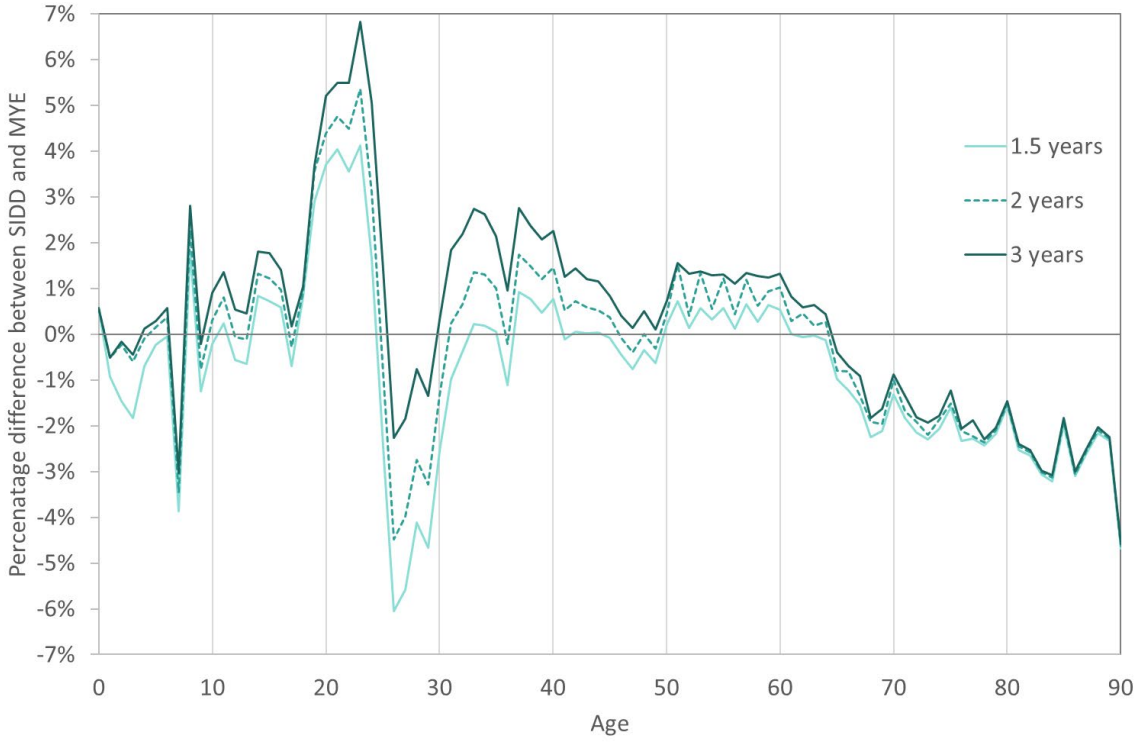


Figure 11 Potential difference between the SIDD and the MYE for females in 2018 for different cut-offs of the Health Activity interaction date.



Notes on statistical publications

Statistical Research

This publication presents statistical research and the methodology is still under development. We welcome any feedback from users on ways in which the methodology or data sources may be developed to improve the quality of these statistics in future years.

National Records of Scotland

We, the National Records of Scotland, are a non-ministerial department of the devolved Scottish Administration. Our aim is to provide relevant and reliable information, analysis and advice that meets the needs of government, business and the people of Scotland. We do this as follows:

Preserving the past – We look after Scotland's national archives so that they are available for current and future generations, and we make available important information for family history.

Recording the present – At our network of local offices, we register births, marriages, civil partnerships, deaths, divorces and adoptions in Scotland.

Informing the future – We are responsible for the Census of Population in Scotland which we use, with other sources of information, to produce statistics on the population and households.

You can get other detailed statistics that we have produced from the Statistics section of our website. Scottish Census statistics are available on the Scotland's Census website.

We also provide information about future publications on our website. If you would like us to tell you about future statistical publications, you can register your interest on the Scottish Government ScotStat website.

You can also follow us on twitter @NatRecordsScot

Enquiries and suggestions

Please get in touch if you need any further information, or have any suggestions for improvement.

Lead Statistician: Lindsay Bennison

Statistics Customer Services telephone: (0131) 314 4299

E-mail: statisticscustomerservices@nrscotland.gov.uk

For media enquiries, please contact: scotlandscensus@nrscotland.gov.uk